EQAO

Education
Quality and
Accountability Office

Français

Students
Parents
Educators
Public
Employment Opportunities

Grade 9 Assessment of Mathematics

Explanatory Paper

The development, scoring and reporting of the Grade 9 Assessment of Mathematics, 2000-2001

This paper provides information relevant to the validity, reliability, comparability and interpretation of EQAO's Grade 9 Assessment of Mathematics, which was administered for the first time in Ontario schools in January and June 2001. We are providing as much information as we currently have available to assist those who are interpreting and using the results. Additional information and reports of a technical nature will be made available on the EQAO Web site as they are completed.

Validity, reliability and comparability are built into all phases of assessment development, scoring and reporting.

Development and Content of the Assessments

The Grade 9 Assessment of Mathematics is based on *The Ontario Curriculum, Grades 9 and 10: Mathematics (1999)*. There were eight separate assessments for students in academic or applied programs, in English or French, in a January or June administration. The assessments at each administration consisted of common items (40%) reflecting curriculum expectations common to both languages and programs, and unique items (60%) reflecting curriculum expectations unique within a language and/or program. Multiple-choice questions, short-answer items and extended response tasks were developed by Ontario teachers to reflect the curriculum strands (Number Sense and Algebra, Relationships, Analytic Geometry, Measurement and Geometry), the categories (Knowledge/Understanding, Problem Solving, Communication, Application) and the general and specific curriculum expectations.

Items on the 2001 assessment were developed and tried out in classrooms by teacher-developers from across the province to refine the tasks and to ensure opportunities for students to demonstrate mathematics skills across all levels of achievement. EQAO education officers reviewed items for

- question specifications (e.g., does it address the right strand, category, number of expectations)
- accessibility (e.g., does the content engage students, is the reading level appropriate, do graphics communicate the necessary information, is the question bias-free)
- mathematical correctness (e.g., is the context mathematically appropriate)
- formatting concerns (e.g., are graphics appropriately sized, placed, are mathematical conventions consistently followed)

Items on the 2001 assessment were piloted with volunteer classes in January 2000, then field tested in May 2000 (volunteers) and January 2001 (random sample of students). At both stages, items were analyzed for difficulty (percent of students scoring at each code) and discrimination (correlation of scores on the item with students' total scores). Based on statistical analyses in conjunction with curriculum coverage requirements, items were selected from those piloted to go forward to field test, and from field test to the full assessment.

Items on the assessments were validated by teachers of mathematics, teachers experienced in equity issues and special needs, university mathematicians and individuals with technical knowledge pertaining to reliability and validity.

The full assessment was constructed to provide multiple opportunities for students to demonstrate mathematics achievement in each curriculum category (a range of 12 to 27 opportunities) and strand (a range of 10 to 28 opportunities), on a variety of item types (multiple-choice, short-answer and tasks), and across all four levels of achievement.

Scoring Guides and Process

The EQAO Reporting Scale is based on the Ontario Achievement Chart and is tailored to large-scale assessment of mathematics. It was used to develop generic descriptors for each category.

The number of descriptors was determined by the complexity of the item. EQAO developed specific scoring guides for each task or item based on the descriptors. Each numerical code on the Specific Scoring Guides identifies a descriptor for the characteristics of the student's work.

In recognition of the various types of assessment items (multiple-choice, short-answer and tasks) the number of codes varied.

- Multiple-choice questions were machine-scored: 0, 1 (wrong or

right).

- Short-answer items were coded by markers: 0, 1, 2.
- Extended-response tasks received multiple scores, for one or more strands and for all categories. They were coded by markers 1, 2, 3, 4, 5 or 6. The number of codes for a particular aspect of a task was dependent on the complexity of the item and students' responses, and the category for which it was being scored.

For short-answer items and extended-response tasks, there were additional codes for student responses that were blank, unintelligible or unrelated to the question.

The Specific Scoring Guides were initially developed theoretically, from mathematics teachers' knowledge of student work. Then they were validated by mathematics' teachers through reviewing student work from the field test. After field test marking, comments from markers were reviewed and further changes were incorporated into the guides.

Prior to marking the full assessment, the Specific Scoring Guides were again validated against student work from the assessment itself, and refined where necessary.

The assessment was marked in July 2001 by approximately 1500 teacher markers. Markers were trained to code students' responses based on the descriptor that best represented the student work. The training included individual coding of training booklets and marking with a partner (paired marking).

Throughout marking, a daily group marking activity was conducted, in which all markers scored the same work. Reports of marker agreement on this activity were used to clarify codes and provide additional assistance to markers where needed. Also, data were analyzed on a daily basis for marking patterns, and booklets were identified for remarking as needed. In addition, a random sample of booklets was re-inserted into the marking flow to provide data for later analysis of marking agreement and reliability.

Analysis and Reporting

Students' scores on the Grade 9 Assessment of Mathematics reflect the number of items for which the student provided scorable responses, the quality of the student's answers as reflected in the scoring codes assigned by the markers, and the characteristics of the items (difficulty, discrimination). Item response theory (IRT) was used to analyze student results, in order to generate the Overall, Category and Strand scores.

IRT Item calibration: Item calibration is the process by which we estimate item characteristics (parameters). Three item parameters are calculated. These are the item discrimination, item difficulty, and a guessing parameter, which is associated only with multiple-choice items. An item's discrimination tells us how well an item distinguishes between students of high achievement and low achievement. An item's difficulty tells us the probability of scoring at a particular code on that item.

Scoring Students' Achievement Using Item Response Theory: After obtaining the item parameters across all students in each language and program group, we scored each student's achievement in each category and strand, and overall. Students' IRT scores thus reflected the number of questions a student answered and the correctness of their responses, and the characteristics of those items. In addition, the IRT scoring model considers how the questions relate to each other and to the student's total performance on the test.

Assessment Comparability and Equating: Within each language and program group, a different form of the same test was administered in January (semester 1 students) and June (semester 2 and full year students). These two assessments were designed to be as similar in content and statistical characteristics as possible. In addition, a set of equating items was administered to a sample of students in both January and June, to form a link between the tests administered on different dates. For each administration, the equating items were calibrated together with the full assessment. The obtained item parameters for the equating items were then used to equate the June assessment to the January assessment. In this way, students' IRT scores were placed on the same scale, making them directly comparable within each language and program group.

Aligning Students Scores with Levels of Achievement: Near the end of the marking process, teachers participated in an item alignment process. Based on their experience with marking, their knowledge of the items and codes, and their knowledge of the Ontario Curriculum Levels of Achievement, markers were asked to make independent judgments of the level or levels of achievement each code represented.

At the end of marking, a number of markers reviewed a random sample of students' work on each task as a whole and assigned the responses on the whole task to a level of achievement. This served as a holistic validation of the mapping of the codes to the Ministry Achievement Chart.

The scoring codes used by the markers and associated levels (from alignment) were combined with item parameters in plots displaying the

probability that a particular code falls in a particular level, at each point of the IRT scale. Please review Scoring the Grade 9 Assessment of Mathematics: Alignment to Levels of Achievement for further information.

Plots were inspected to locate the "cutpoints" between levels where the probability lines intersect. The accuracy of assignment of students to each level of achievement was maximized, and the distribution of students' scores and markers' judgments were represented fairly. Raw score distributions and the sample of holistically scored student work were used to assist in this process. Boundaries (i.e. cutpoints) were established separately for each language and program, and for each type of score to be reported, i.e., overall scores as well as category and strand scores. The resulting sets of boundaries were again inspected for consistency and consolidated or refined. They were then applied to all students' IRT scores and interpolated within each level of achievement to convert them to a reporting score. Since students' IRT scores were based on work completed on the assessment, the reporting scores were also adjusted for the proportion of the full assessment the student missed, either due to absence or skipped items. Reliable reporting scores could not be provided for students who missed/skipped more than half of the assessment. These students were deemed to have provided "Insufficient Information to Score" (IIS).

Score Variability and Reliability: Students' performance may fluctuate somewhat from day to day or from one set of tasks to another, for a variety of reasons. On the Individual Student Report, the shaded range around a student's overall score represents an estimate of the amount by which a student's score could vary in relation to other students scoring at the same level. These ranges may also be useful in consideration of students' strengths and areas for improvement among categories and strands, in relation to their overall score. Programming needs for individual students or groups of students can also be identified in relation to the provincial standard.

Education Quality and Accountability Office, Suite 1200, 2 Carlton Street, Toronto, ON M5B 2M9

Telephone: 1-888-327-7377 • Fax: 416-325-2956

Please send your comments and general enquiries to E-mail@eqao.com
Questions or comments regarding EQAO's Web site can be sent to webmaster@eqao.com

Certain publications on this site are provided as Adobe Acrobat PDF files. You can download this free software from the Adobe Web site.

© 2003, Queen's Printer for Ontario

This page was last updated on October 1, 2003.

Office de la qualité
et de la responsabilité
en éducation

English

Élèves
Parents
Éducatrices et éducateurs
Public
Occasions d'emploi

Test de mathématiques, 9e année

Rapport explicatif

L'élaboration, la notation et la communication des résultats du Test de mathématiques, 9e année, 2000-2001

Ce rapport présente des renseignements relatifs à la validité, la fidélité, la comparabilité et l'interprétation du Test de mathématiques, 9e année de l'OQRE. Ce test a été administré pour la première fois dans les écoles de l'Ontario en janvier et en juin 2001. Le Rapport explicatif fournit des renseignements qui sont présentement disponibles pour aider les personnes qui désirent interpréter et utiliser les résultats. Des informations supplémentaires, ainsi que des rapports techniques, seront téléchargés dans le site Web de l'OQRE, au www.eqao.com, dès qu'ils seront disponibles.

La validité, la fidélité et la comparabilité sont comprises dans toutes les étapes de l'évaluation, de la notation et de la communication des résultats.

Élaboration et contenu des évaluations

Le Test de mathématiques, 9e année est fondé sur *Le curriculum de l'Ontario, 9e et 10e année – Mathématiques, 1999*. Il comprend huit versions distinctes, en français et en anglais, qui ont été administrées aux élèves inscrit(e)s à un cours théorique ou appliqué en janvier et en juin 2001. Chaque test comprend des items communs (40 %) qui sont liés aux attentes communes du programme-cadre de mathématiques de langue française et de langue anglaise, des items uniques (60 %) qui sont liés aux attentes de chaque programme-cadre de mathématiques et aux cours théorique et appliqué. Les questions à réponse choisie, les questions à réponse courte et les tâches à réponse élaborée ont été rédigées par une équipe d'enseignantes et d'enseignants afin de refléter les domaines mathématiques (Numération et algèbre, Relations, Géométrie analytique et Mesure et géométrie), les compétences (Connaissance et compréhension, Résolution de problèmes, Communication, Mise en application) et les attentes et les contenus d'apprentissage du programme-

cadre de mathématiques.

Les items du test de janvier et de juin 2001 ont été élaborés et mis à l'essai dans les salles de classes de toutes les régions de la province, par des enseignantes et enseignants, afin de s'assurer que les élèves puissent faire les tâches d'évaluation et démontrer leurs habiletés en mathématiques à tous les niveaux de rendement. Les agentes et agents d'éducation de l'OQRE ont révisé les items selon les critères suivants :

- La spécificité de la question (c.-à-d., Évalue-t-elle le domaine et la compétence appropriés?);
- L'accessibilité (c.-à-d., Est-ce que le contenu est stimulant pour l'élève? Est-ce que le niveau de difficulté du texte à lire est approprié? Est-ce que les graphiques communiquent bien l'information demandée? Est-ce que la question est exempte de préjugés?);
- L'exactitude mathématique (c.-à-d., Est-ce que le contenu mathématique est approprié?);
- La mise en page (c.-à-d., Est-ce que la grosseur et l'emplacement des graphiques sont appropriés? Est-ce que les conventions mathématiques sont systématiquement respectées?).

Les items des tests de 2001 ont été mis à l'essai à trois reprises : 1) une mini mise à l'essai auprès de classes volontaires a eu lieu en janvier 2000; 2) une mise à l'essai auprès de classes volontaires a eu lieu en mai 2000; 3) une mise à l'essai auprès d'un échantillon aléatoire d'élèves a eu lieu en janvier 2001. À chaque étape, les items ont été analysés pour leur niveau de difficulté (pourcentage d'élèves dont la note se situe à chaque niveau) et la corrélation (comment les notes sur les questions individuelles se rapportent aux notes totales des élèves). Les items ont été sélectionnés selon les données statistiques en conjonction avec les exigences du curriculum. Dans un premier temps, des items de la mini mise à l'essai ont été sélectionnés pour l'étape suivante. Dans un deuxième temps, des items faisant partie de la mise à l'essai ont été sélectionnés pour être inclus dans l'évaluation complète.

Les items des tests ont été validés par des enseignantes et enseignants en mathématiques, des expertes et experts en matière d'équité et des besoins particuliers, des professeures et professeurs en mathématiques de niveau universitaire, des personnes ayant des connaissances techniques liées à la validité et la fidélité des évaluations.

Le test a été élaboré afin de donner aux élèves de nombreuses occasions pour faire preuve de leurs connaissances et de leurs habiletés en mathématiques dans chaque compétence du programme-cadre (gamme

de 12 à 27 occasions) et dans chaque domaine (gamme de 10 à 28 occasions) selon une variété de types d'items (questions à réponse choisie, à réponse courte et tâches) et selon les quatre niveaux de rendement.

Guides d'encodage et processus

L'échelle de rendement de l'OQRE est fondée sur la grille d'évaluation de rendement du ministère de l'Éducation et est adaptée pour une évaluation de mathématiques à grande échelle. Elle a été utilisée pour élaborer des codes descripteurs génériques pour chaque compétence.

Le nombre de codes descripteurs a été déterminé selon la complexité de chaque item. L'OQRE a élaboré des guides d'encodage spécifiques pour chaque tâche ou item en fonction des codes descripteurs. Chaque code numérique, sur les guides d'encodage spécifiques, représente un code descripteur identifiant les caractéristiques du travail de l'élève.

Selon les différents types d'items du test (questions à réponse choisie, questions à réponse courte et tâches à réponse élaborée), le nombre de codes varie.

- Les questions à réponse choisie ont été notées électroniquement à l'aide d'un système d'encodage 0, 1 (correct ou incorrect).
- Les questions à réponse courte ont été codées par les correctrices et correcteurs en utilisant 0, 1, 2.
- Les tâches à réponse élaborée ont reçu différentes notes pour un ou plusieurs domaines et pour toutes les compétences. Les correctrices et correcteurs ont utilisé un système de codes 1, 2, 3, 4, 5 ou 6. Le nombre de codes, pour un aspect particulier de la tâche, dépendait de la complexité de celle-ci, des réponses de l'élève ainsi que de la compétence pour laquelle elle était notée.

Pour les questions à réponse courte et les tâches à réponse élaborée, il y avait des codes supplémentaires pour les questions qui n'avaient pas été répondues, qui étaient illisibles ou dont les réponses n'étaient pas reliées à la question.

Au départ, les guides d'encodage spécifiques ont été élaborés suivant une approche théorique par des enseignantes et enseignants de mathématiques. Ils ont ensuite été validés par les enseignantes et enseignants de mathématiques qui avaient révisé les travaux des élèves tirés de la mise à l'essai. Après la notation de la mise à l'essai, les commentaires des correctrices et correcteurs ont été pris en considération et d'autres modifications ont été apportées aux guides.

Avant la notation de l'évaluation complète, les guides d'encodage spécifiques ont été validés de nouveau à partir de travaux des élèves tirés du test et ont été mis au point lorsque nécessaire.

Le test a été noté en juillet 2001 par environ 1 500 enseignantes et enseignants. Ces personnes ont reçu une formation pour utiliser un système de codes pour noter les réponses des élèves. Ces codes ont été élaborés à partir de codes descripteurs qui étaient le plus représentatif du rendement de l'élève. La formation comprenait également l'encodage de cahiers de formation ainsi que la notation avec partenaire.

À chaque jour, une activité de groupe avait lieu. Lors de cette activité, l'ensemble des correctrices et correcteurs notaient le même travail de l'élève. Les rapports d'entente des correctrices et correcteurs servaient de guide pour vérifier les codes ou pour offrir une aide supplémentaire aux personnes qui en avaient besoin. Les données étaient analysées quotidiennement pour identifier les façons de procéder et certains cahiers étaient identifiés pour qu'ils soient notés une deuxième fois, lorsque nécessaire. De plus, un échantillon aléatoire de cahiers ont été réinsérés dans la pile de cahiers à noter afin de fournir des données pour des analyses subséquentes sur les coefficients d'entente et la fidélité du processus.

Analyse et communication des résultats

Le niveau général de rendement des élèves au Test de mathématiques, 9^{e} année représente le nombre d'items auxquels les élèves ont fourni des réponses pouvant être notées, la qualité des réponses des élèves telle que représentée dans le système de codes utilisé par les correctrices ou correcteurs et les caractéristiques des items (difficulté, corrélation). La Théorie de la réponse à l'item (TRI) a été utilisée pour analyser les résultats des élèves afin de générer le niveau général de rendement pour les composantes du test ainsi que pour les domaines et les compétences.

Calibrage des items selon la Théorie de réponse à l'item (TRI) : Le calibrage des items est un processus qui permet d'estimer les caractéristiques des items (paramètres). Trois paramètres sont calculés : le pouvoir discriminant de l'item, la difficulté de l'item et le paramètre de pseudo-chance (qui est utilisé seulement pour les questions à réponse choisie). Le pouvoir discriminant d'un item nous indique dans quelle mesure l'item nous permet de distinguer entre les élèves qui ont un rendement élevé et les élèves qui ont un rendement faible. La difficulté d'un item nous indique la probabilité d'avoir un code particulier pour cet item.

La notation du rendement des élèves selon la Théorie de réponse à l'item (TRI) : Après avoir obtenu les paramètres de l'item à partir des résultats de l'ensemble des élèves de langue française et de langue anglaise inscrits dans un cours théorique ou appliqué, nous avons établi le niveau de rendement de chaque élève pour chaque domaine, pour chaque compétence et pour toutes les tâches d'évaluation. La TRI des élèves reflétait ainsi le nombre d'items auxquels l'élève avait répondu, l'exactitude des réponses ainsi que les caractéristiques de ces items. De plus, le modèle de la TRI tient compte de la façon dont les questions sont reliées entre elles et du rendement global de l'élève au test.

Comparabilité des évaluations et mise en équivalence : Pour le cours théorique et appliqué, en français et en anglais, une version différente du test a été administrée en janvier (pour les élèves du premier semestre) et en juin (pour les élèves du deuxième semestre et des cours annuels). Ces deux tests ont été élaborés en parallèle pour que le contenu et les caractéristiques statistiques puissent être comparés. En plus, un ensemble d'items de calibrage a été administré à un échantillon aléatoire d'élèves, au mois de janvier et au mois de juin, pour établir un lien entre ces tests. À chaque administration, la mise en équivalence des items a été faite lors du calibrage de l'évaluation complète. Les paramètres obtenus pour les items de calibrage ont ensuite été utilisés pour le test administré en juin par rapport à celui administré en janvier. Les notes de la TRI des élèves ont été inscrites sur la même grille, les rendant ainsi comparables pour le cours théorique et appliqué de langue française ou de langue anglaise.

Alignement des résultats des élèves avec les niveaux de rendement

Alignement des items : Vers la fin du processus de notation, les enseignantes et enseignants ont participé au processus d'alignement des items. Selon leur expérience de notation, leur connaissance des items et du système de codes, ainsi que leur connaissance des niveaux de rendement du Curriculum de l'Ontario, les correctrices et correcteurs ont porté des jugements indépendants sur le niveau ou les niveaux de rendement que chaque code représente.

Alignement holistique : À la fin de la notation, quelques correctrices et correcteurs ont révisé un échantillon aléatoire de travaux d'élèves pour chaque tâche et ont attribué un niveau de rendement à toutes les réponses de la tâche. Ce procédé a servi comme validation holistique de l'application des codes de la grille d'évaluation du rendement du ministère de l'Éducation.

Les codes de notation utilisés par les correctrices et correcteurs et les niveaux correspondant (à partir de l'alignement) ont été combinés aux

paramètres des items sous forme de graphiques montrant la probabilité qu'un code donné soit associé à un niveau donné, pour chaque point de l'échelle de la TRI.Veuillez vous référer au document La notation du Test de mathématiques, 9e année: Alignement des niveaux de rendement pour de plus amples informations.

Les graphiques ont été examinés afin de situer les bornes entre les niveaux correspondants à l'intersection des courbes de probabilité. De cette façon, l'exactitude avec laquelle un niveau de rendement a été attribué à chaque élève a été maximisé et la distribution des résultats des élèves et du jugement des correctrices et correcteurs ont été représentés de façon juste. La distribution des résultats bruts et l'échantillon des travaux d'élèves notés de façon holistique ont été utilisés dans le processus. Les bornes ont été établies séparément pour le cours théorique ou appliqué en français et en anglais et pour chaque type de résultat faisant l'objet d'un rapport, c'est-à-dire le résultat général, le résultat pour chaque compétence et pour chaque domaine. L'ensemble des bornes que l'on a obtenu a été de nouveau examiné afin d'en vérifier la cohérence, pour les combiner ou les raffiner. Elles ont ensuite été appliquées à tous les résultats de la TRI des élèves et interpolées dans chaque niveau de rendement pour les convertir en un résultat final. Puisque les résultats de la TRI des élèves étaient basés sur le travail effectué lors de l'évaluation, les résultats finaux ont également été ajustés pour compenser pour la proportion de l'évaluation finale non complétée par l'élève en raison d'une absence ou d'items ignorés. Pour les élèves ayant manqué ou ignoré plus de la moitié du test, aucun résultat fidèle n'a pu être fourni. La mention « Informations insuffisantes pour être notées (I. I.) » a été attribuée à ces élèves.

Variabilité des résultats et fidélité

La performance des élèves peut varier d'un jour à l'autre, ainsi que d'un ensemble de tâches à l'autre, pour de multiples raisons. Dans le Rapport individuel de l'élève, la zone ombrée autour du résultat global de l'élève représente une estimation du degré de variation possible du résultat de l'élève en relation avec les autres élèves ayant un résultat au même niveau de rendement. Ces zones peuvent également être utiles pour prendre conscience des forces et des faiblesses de l'élève et pour déterminer les aspects à améliorer au niveau des compétences et domaines selon leur résultat global. La programmation des besoins pour les élèves ou groupes d'élèves peut également se faire par rapport à la norme provinciale.

Office de la qualité et de la responsabilité en éducation, 2, rue Carlton, bureau 1200, Toronto (Ontario) M5B 2M9

Téléphone : 1 888 327-7377 • Télécopieur : 416 325-2956

Commentaires et renseignements généraux : e-mail@eqao.com
Veuillez envoyer vos questions ou commentaires concernant le site Web à : webmaster@eqao.com

Certaines publications téléchargées dans ce site sont offertes en version Adobe Acrobat PDF.
Vous pouvez télécharger ce logiciel gratuit à partir du site Web d'Adobe.

© Imprimeur de la Reine pour l'Ontario, 2003

Dernière mise à jour le 1er octobre 2003